BIOWHERE – GEOREFERENCING NEW ZEALAND'S BIOTA FROM TEXTS

Kalana Wijegunarathna[1], Kristin Stock[1], Christopher Jones[2], Aron Wilton[3], Jonathan Procter[4], Hone Morris[5], David Medyckyj-Scott[3], Fraser Morgan[3], John Wieczorek[6], Brandon Whitehead[3]

[1]Massey Geoinformatics Collaboratory – Massey University, New Zealand. kalanainduwara.16@cse.mrt.ac.lk, k.stock@massey.ac.nz

[2]School of Computer Science and Informatics – Cardiff University, United Kingdom. jonescb2@cardiff.ac.uk

[3]Manaaki Whenua Landcare Research. wiltona@landcareresearch.co.nz, medyckyj-scottd@landcareresearch.co.nz, morganf@landcareresearch.co.nz, whiteheadb@landcareresearch.co.nz

[4]School of Agriculture and Environment – Massey University, New Zealand. j.n.procter@massey.ac.nz

[5]Te Pūtahi-a-Toi. h.w.Morris@massey.ac.nz

[6]Rauthiflor LLC. tuco@berkeley.edu

Abstract

BioWhere aims to develop techniques that can be harnessed to map large volumes of biota specimens from all over New Zealand and Antarctica. With over 12 million records, these specimens along with their locations are georeferenced textually, usually with complex natural language descriptions, across various scientific publications and specimen collections held by museums and other institutions. However, only a small fraction of these specimens has been mapped owing to the large volume of the data and the amount of labour the process of converting textual descriptions into coordinates demands. The challenge is amplified further due to the indefinite, vague nature of natural language used to textually describe the locations of these specimens. The automated tools currently in use fall short of effectively mapping the specimens because they rely on incomplete gazetteers and ignore spatial language.

The BioWhere project will explore the latest techniques in natural language processing and develop methods that can effectively overcome the aforementioned shortcomings using computational georeferencing methods to translate human language descriptions of locations to geographical coordinates. Our machine learning models will incorporate environmental factors, linguistic context, and the characteristics of the named place in building a self-learning gazetteer. Incorporating physical, historical, and cultural context will enrich the gazetteer with Māori knowledge including the origin, narrative and meaning of Māori place names. This will unlock vast amounts of structured scientific knowledge that are currently inaccessible. Furthermore, methods developed in the project will find further applications in a range of domains including disaster response, cultural heritage, and health.

Keywords

Natural Language Processing, Data Science, Biodiversity, Georeferencing, Geospatial Natural Language, Geographic Information Retrieval.